International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 7 (2015) pp. 106–114
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Exploring Association Rules in a Large Growing Knowledge Base

Rafael Garcia Leoenl Miani[1] **and Estevam Rafael Hruschka Junior**[2]

[1]Federal Institute of Sao Paulo - IFSP, Informatic Department,
Jerôimo Figueira da Costa 3014,Votuporanga 15503-110 , Brazil
rafagami@gmail.com

[2]Federal University of Sao Carlos - UFSCar, Computer Department,
Washington Luis 235, Sao Carlos 13565-905, Brazil
estevam@dc.ufscar.br

Abstract: **Large growing knowledge bases have been an interesting field in many researches in recent years. Most techniques focus on building algorithms to help the Knowledge Base (KB) automatically (or semi-automatically) extends. In this article, we make use of an association (or generalized association) rule mining algorithm in order to populate the KB and to increase the relations between KB's categories. Considering that most systems constructing their large knowledge bases continuously grow, they do not contain all facts for each category, resulting in a missing value dataset. To accomplish that, we developed a new parameter, called MSC (Modified Support Calculation) measure. This measure also contributes to generate new and significant rules. Nevertheless, association rules algorithms generates many rules and evaluate each one is a hard step. So, we also developed a structure, based on pruning obvious itemsets and generalized association rules, which decreases the amount of discovered rules. The use of generalized association rules contributes to their reduction. Experiments confirm that our approaches discover relevant rules that helps to populate our knowledge base with instances (by MSC measure and association rules), increase the relationships between the KB's domains (using generalized association rules) as well as facilitate the process of evaluating extracted rules (pruning obvious itemset and association rules).**

Keywords: Association rules, large knowledge base, missing values, obvious rules, obvious itemset

I. Introduction

Large growing knowledge bases have been widely explored in many works in past few years [1]. Cyc [2], DBpedia[3], NELL [4] and YAGO [5] are some systems that were constructed aiming to build these knowledge bases.

In order to extend their knowledge base (KB), many different techniques were developed. Some with the objective to extend their instances, by populating the KB with them. But others approaches try to increase the relations between the knowledge base categories. In this way, we developed a technique, based on an association rule mining algorithm to both populate the NELL's KB with instances and to increase the relations between NELL's categories.

NELL (Never-Ending Language Learning) is a computer system that runs 24 hours per day, 7 days per week, extracting information from web text to populate and extend its own KB. The main goal of the system is to learn to read the web better each day and to store the gathered knowledge in a never-ending growing KB. The system takes advantage of many different components like CPL [6], CSEAL [7], Prophet [8], OntExt [9], and Conversing Learning [10], in order to be self-supervised and avoid semantic drifting [11]. NELL's knowledge base is represented by an ontology-based structure characterized by categories, relations and their instances.

Since NELL's KB continuously grows each day, it does not contain all instances of every category, neither all instances of every relation described in the ontology. Also, some relations may never occur for some specific instances. So, NELL's KB contains many missing values. To extract rules from a knowledge base having such missing values properties, we have two main options: i) exploring new approaches when using frequent-item-set-based algorithms; ii) trying relational rule extraction approaches (as the one presented in [12], for instance). In this work, we focus on the first possibility. The work described in this paper shows how we can help NELL populating its own knowledge base using association rule mining algorithms [13] and how we can increase the relations between its categories by using Generalized Association Rules (GARs)[14]. We aim at using facts (knowledge) already stored in NELL's KB as input to a frequent-item-set-based algorithm, thus, discovering new relations and instances that NELL's algorithms were not able to get from web text before as well as increasing it with interesting relations brought by GARs. The proposed approach can make use of any association rule mining algorithm, but in the experiments described in this paper we have used NARFO algorithm [15], wich will be better described in methodology section.

There are two main reasons that explain the presence of missing values in a large and continuous growing knowledge base like NELL. First, some instances were not extracted yet from the web by NELL's algorithms, i.e., they are not available at the moment (but should be in the future). Second, a specific relation between two categories might be defined in the ontology, but may never occur in web text, thus, no instances will be extracted by NELL's algorithms. This will be explained later. In this context, we modified traditional support measure, creating an adapted parameter called MSC (Modified Support Calculation).

We also have to analyze extracted rules to verify which one is a valid rule that can be used to help to populate the KB. Our technique differs from traditional association rule algorithms on how to interpret and weight the extracted rules. Usually, a rule is considered strong if both support and confidence measures are greater than a threshold, and the higher support and confidence of a rule, the stronger the rule. As NELL's ontology grows each day, daily learning just a small portion of what can be learned, a rule with high confidence must be carefully analyzed, as it could bring wrong knowledge, as NELL does not have all instances to every category yet. Therefore, those rules (with high confidence) need to be carefully inspected before being considered useful. MSC measure and the process of analyzing each rule extracted that can help populating the KB grows was validated in [16].

Nevertheless, this paper also describe the use of a Generalized Association Rule (GAR) mining algorithm [14], in order to extend the relations between NELL's KB domains. In this way, we expect that the use of generalized association rules can bring new relations between the categories of the KB.

Beyond that, the use of GARs can be an effective technique to help the process of analyzing the discovered rules. Using this approach the amount of extracted rules must decrease significantly depending on the domain's characteristic, once specifics association rules can be represented by their generalized ones. For example, imagine that the generalized association rule *personWinAwardTrophyTournament(X, Y) → personPlaysSport(X, Z)* is discovered. If this rule was not generated, we have to evaluate too many specifics association rules as there are many people who win different kind of tournaments in many sports. With GAR we might be able to increase the KB by increasing the relations between these domains that were not discovered yet by NELL's algorithms.

To improve the extracted rules evaluation we also explored the treatment of obvious association rules and itemsets. Applying that, the effort of evaluating the discovered rules may be, each time, reduced, once the obvious discovered rules must not be showed in futures iterations of the algorithm. Note that the environment the algorithm takes place is a never ending growing knowledge base. So, without an obvious rule and itemset treatment, the process of analyzing each rule gets continuously harder.

We consider a rule as an obvious rule if it was already discovered before and is either relevant or not. By relevant rule, we mean association or generalized association rules already discovered that helped the KB grows by populating it with new instances or by extending the relations between categories. If a rule is significant or not, it is inserted in a table as we are not interesting in analyze them again in futures algorithm's iterations. This method will be better explained later.

The process of increasing KB's relations by using GARs and the approach of obvious itemsets and association rules treatment was described in [17].

In this way, this article aims to describe how all the following contributions work combined:

1. How association rules may help populating a large KB;
2. How the new MSC measure works in a large growing knowledge base consisted of many missing values;
3. How generalized association rules may help the discover of new relations in a large KB;
4. How effective and necessary is to analyze each extracted association rule discovered;
5. How the treatment of obvious discovered association and generalized association rules could help improving the process of analyze rules in futures iterations.
6. How the use of generalized association rules minimizes the evaluation of the discovered rules.

II. Related Work

Works exploring large growing knowledge bases are rising in recent years. BabelNet [18], Cyc [2], DBpedia [3], NELL [4], ReVerb [19], YAGO [5] and YAGO2 [20], which extend YAGO to deal with spatio-temporal dimension, are systems that have been developed techniques to help their KBs extends.

Most works focus on increasing the knowledge base by populating it with new instances or by enlarging the relations between their categories or domains. In order to populate the KB, [21] presents EVIN, a system that extracts events from news articles and organizes them into semantic classes to populate a KB. [22] proposed a knowledge base population methodology in which they consider temporal data in the process. An approach that automatically extracts entities and relationships from textual documents were developed in [23]. They extracted entities and their relationships are verified using classification and linking evidence based techniques. Other related efforts in this field can be found in [1],[24], [25], [8] and [9].

In this work, we propose the use of association rules in order populate the knowledge base with instances and GARs to investigate how useful they can be to extend the relations between the knowledge base's categories. NARFO algorithm [15], which is a generalized association rule mining algorithm, is used to discover new rules. NARFO can navigates through KB's instances and categories, producing generalized association rules. It also has a redundant treatment to verify if a specific rule is included in a generalized one. This algorithm introduced a new parameter, called *mingen* (minimal generalization) [26] to generalize association rules.

Similar works that use association rule mining to extend ontologies or KBs were developed in [27] and [1]. The first one uses association rules under an existing ontology, the Finnish General Upper Ontology YSO [28], in order to improve it by populating more relationships between its concepts. [1] produced AMIE, an association rule mining algorithm under incomplete evidence. AMIE focuses on discovering relations like if motherOf(m; c) $\wedge$ marriedTo(m; f) $\rightarrow$ fatherOf (f; c). PROSPERA [29], which is based on SOFIE framework [30], is another approach that uses an association rule algorithm to discover relations from Web Text in order to populate their KB.

In a system like NELL, the KB is consisted by many missing values. As stated in [31], many association rules mining algorithm are used based on databases that do not contain missing values. To accomplish this field, we modified traditional support calculation, creating a new measure, called MSC [16]. Basically, it discards an itemset if all items in the specific itemset are missing.

According to [32], there are some reasons that explain why empty cells occur in datasets:

1. Values recorded are missing because they were too small or too large to be measured;

2. Values recorded are missing because they have been forgotten, they have been lost or they were not available.

There are some strategies to deal with the problem of missing values for association rule mining. In [31], the authors apply a technique in which an itemset with missing values is not considered, i. e., it is disabled. For that, they created a new measure definition to support calculation. In [33], all items corresponding to attributes with missing values are set to zero. An algorithm (called AR) was proposed in [31], in which missing values are replaced by a probability distribution over possible values represented by existing data. XMiner [34] is an association rule algorithm based on [27] that also brings the concept of extensible itemsets to deal with missing values.

The works presented in [35], [36] both focused on the problem of missing values for association rule mining. In [35], a Markov-chain based Missing Value Estimation (MC-MVE) method was proposed. The work proposed in [36] developed an iterative way to extract association rules for inferring missing values based on the algorithm created by [31].

In the context of NELL's ontology, missing values occur either because they are not available at the moment, or because the relation may never happen. NELL's algorithms extract relations between categories, like *athleteplaysport (X,Y)* and *athleteplaytournament (X, Z)*. So, by transitivity, we can induce that if an athlete X plays a sport Y and if the same athlete X plays a tournament Z, then this tournament (Z) is related to that sport (Y). However, NELL could extract relations like *athleteplaysport (Lebron_James, basketball)*, *athleteplaytournament (kevin_garnett, nba)* and *athleteplaytournament (Lebron_James,nba)*. Notice that, in this given example, NELL did not extract the relation between kevin_garnett and the sport he plays and, thus, this value will be missing in the KB (kevin_garnett is also a basketball player). Another kind of missing value that NELL's KB has is when the relation between two categories did not occur and will never happen. Following along these lines, this paper focuses on dealing with two kind of missing values:

1. Missing values that are missing because they are not available yet;

2. Missing values caused by relations between two specific categories that may never happen.

An important problem on the process of extracting association rules is how evaluate the amount of rules extracted. Therefore, using GARs might reduce the quantity of rules and facilitate the evaluating process. NARFO, as mentioned before, has a generalization process and a redundancy treatment that helps reducing the number of extracted rules.

Most works on this issue propose post-processing techniques to prune the number of rules. In [37], a post-processing task is developed to decrease the amount of association rules. They propose an interactive approach where human domain experts filter the rules extracted. CoGAR [38] is a GAR algorithm that introduced two new measures: (i) a schema constraint is created by an analyst and drives the itemset mining phase and (ii) opportunistic confidence constraint that identifies significant and redundant rules at post-processing phase. PNAR_IMLMS [39] and MIPNAR_GA [40] are algorithms that discovery both positive itemsets (frequent itemsets) and negative itemsets (infrequent itemsets). They created some measures to prune rules and generate positive and negative association rules.

Nevertheless, many efforts have been taken on checking if rules discovered before should be displayed again in the future. We call these types of rule as obvious rules. TOPSIL-Miner [41] is an algorithm that store potential significant itemsets in a structure called TOPSIL-Tree and then prune trivial nodes from the current dataset. OPUS_AR [42] reduced the search space adding some constraints on relationships between association rules in order to decrease the number of rules.

In this article, we use four of the above described techniques to reduce the amount of discovered rules:

1. Eliminating obvious or trivial itemsets during the candidate's generation;

2. Eliminating obvious or trivial generalized association rules in post-processing;

3. Use generalized association rules and;

4. Perform redundancy treatment at post-processing step.

Besides that, we also have an important contribution: how the use of generalized association rules can increase the relationships between the categories of a large growing knowledge base. Conversing Learning [10], a NELL's component, is used to validate the rules.

Table 1 brings us a comparison between our approach and some related work described in this section.

III. Methodology

The approach described in this article has the following main purposes:

1. Verify how association rules can help populating NELL's KB;

2. Verify how generalized association rules can help increasing the relations between the categories of NELL's KB;

3. Create MSC measure to deal with missing values;

4. Eliminate obvious itemsets in the candidate generation step.

Table 1: Comparison among approaches

Method / Algorithm	Missing values	Post-processing Treatment	GAR	Redundant Treatment	Obvious Rules	Obvious itemsets	Large KBs
AR	X						X
XMiner	X						X
AMIE							X
PROSPERA							X
CoGar		X	X	X			
PNAR_IMLMS		X		X			
MIPNAR_GA		X		X			
TOPSIL-Miner				X		X	
OPUS_AR		X		X			
OUR APPROACH	X	X	X	X	X	X	X

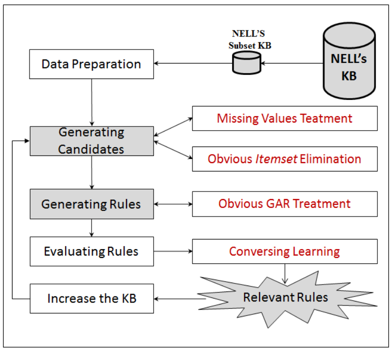

Figure. 1: Steps of the algorithm

5. Prune obvious generalized association rules in post-processing step;
6. Reduce the amount of generated rules by a generalization and redundancy treatment.

In this section, we detail each relevant step of our method. Figure 1 depicts the steps.

A. Data preparation

In this step, the data to be used by NARFO algorithm is prepared. Basically, NARFO is an association rule mining algorithm that also generalize association rules when necessary, that navigates through the ontology structure to be able to identify the instances and their domains. It also has a new parameter called *minGen*, which generalizes an item of a rule if *x%* of a domain's items appears in rules with same antecedents / descendants. A small piece of NELL's KB is selected and an ontology, based on their instances and categories is created.

B. Generating Candidates

This section describes two of our contributions: (i) missing values treatment and (ii) obvious itemset elimination.

1) Missing Values Treatment

A large knowledge base like NELL, as aforementioned, contains many empty cells. To perform an association rule mining algorithm in such environment, we created a new parameter, called MSC (Modified Support Calculation) measure to deal with missing values. To sum up, this technique discard an itemset during support counting if all items on it are missing.

Consider the toy dataset shown in Table II (a), which contains instances of three different categories (*SportsLeague, Sport and Trophy*). Empty cells in Table II (a) represent missing values. So, when calculating MSC, the algorithm does not count missing values to calculate an itemset support value. In Table II (a), *Sportsleague* category has a single empty cell, and this is not considered during support calculation. *Sport* and *trophy* categories have, respectively, four and three empty cells. Table II (b) presents itemsets support values using MSC measure.

MSC measure computation is as follows: for an *n-itemset*, the algorithm discards a missing value if *n items* in the *itemset* are not present. For example, in *1-itemset*, if one missing value is found, the instance containing the missing value is not considered to support calculation. For a *2-itemset*, both items in the *itemset* have to be missing to discard them. The algorithm continues performing this task until no more *itemset* can be found.

For each *itemset* found, the algorithm gets the number of rows in which all items in the current *itemset* appear. Then, it checks the number of rows containing all missing values in the domains of the present *itemset*. Finally, it calculates the support of the *itemset*.

2) Obvious Itemset Elimination

```
Algorithm 1 obiviousItemsetElimination Algorithm
for $i = 0$ to $candidateListSize - 1$ do
    $candidateItemset = candidateList(i);$
    $control = false;$
    for $j = 0$ to $obviousDiscoveredItemsetSize$ do
        if $candidateItemset = obviousDiscoveredItemset(j)$ then
            $control = true;$
        end if
    end for
    if $!control$ then
        $frequentItemset.add(candidateItemset);$
    end if
end for
```

The candidate generation step is very similar to Apriori and other association rules mining algorithms. However, we check each candidate itemset and compare them with a table containing obvious itemsets. If there is a match, the itemset is pruned and will not be considered anymore to generate association rules. Algorithm1 describes this procedure. By an obvious itemset, we mean one that is already used before in a

Table 2: Example to support calculation

(a) dataset sample

SportsLeague	Sport	Trophy
Nba	Basketball	
Nba		nba_championship
		nba_championship
Nba		nba_championship
Nba	Basketball	nba_championship
Nhl		
Nhl	Hockey	

(b) itemsets' support

Itemset	MSC-based Support value
Nba	4/6 = 0.667
Nhl	2/6 = 0.333
Basketball	2/3 = 0.667
Hockey	1/3 = 0.333
nba_championship	4/4 = 1
Nba, basketball	2/6 = 0.333
Nhl, hockey	1/6 = 0.167
Nba, nba_championship	3/7 = 0.428
Nba, basketball, nba_championship	1/7 = 0.142

discovered association rule. It is important to note that if the rule is significant it is inserted in the table (we already discovered it, why show it again in futures iterations?). Although, if a rule discovered is not interesting, it is also inserted in the table, once we do not want that rule showing up again. Only rules that Conversing Learning not conclude yet if is relevant or not are not inserted in the table.

C. Generating Rules

This step focuses on the following issues:

1. Generate association rules;
2. Check for obvious generalized association rules;
3. Generalize rules;
4. Eliminate redundant rules.

The novelty in this step is the generalization and redundancy treatment, which are different comparing to other approaches. NARFO deals with the problem of generalized association rules and also has a redundancy treatment. To generalize the extracted rules, it created a new measure called *mingen*. This generalizes an item of a rule if x% of its domain's instances is presented in other rules with same antecedent(s) or descendant(s).

Also, we added to its implementation an obvious generalized rules treatment. The reason we only take care to obvious generalized rules in this step is that obvious specifics itemsets were already pruned out of the frequent itemset during the candidate generation. The algorithm is pretty much the same as Algorithm 1, but instead of prune of obvious itemsets, it prunes trivial generalized rules. So, only new interesting patterns will be displayed. In this way, the effort of evaluating each rule is reduced.

D. Evaluating Rules

Traditional association rule mining algorithms consider a rule strong that one with support and confidence values higher than the minimum support and confidence values defined. So, those algorithms display the strong association rules extracted. But this fact does not apply in large growing knowledge bases like NELL. We noticed that some rules that have high confidence, specially, could not be good examples of rules to populate and grow the KB, since NELL's KB did not contain all instances and relations yet. And this is one important contribution in this work.

Consider Table III as an example. Imagine that NELL's algorithms extracted some instances and relations about *Players*, *Sports* and *Sportsleagues* as shown in Table III. Based on those instances, it is possible to conclude that a traditional association rule mining algorithm would probably extract the following rule: if *playerPlaysSport (X, Soccer)* then *playerPlaysChampionship (X, italian_championship)*. Notice that, based on Table III, this rule may have high support and confidence, but it carries wrong information. Why does it occur? It does, mainly because NELL's KB does not have all instances and relations yet. So, an important task is to analyze the discovered association rules and use only the useful ones. To analyze each non obvious rule extracted in previous section, we make use of the NELL's Conversing Learning component. To sum up, this component uses Twitter and Yahoo Answers to validate rules. In this work, however, only Twitter was used to evaluate extracted rules. With Conversing Learning help, we intend to automatically evaluate extracted rules and increase NELL's KB with the useful ones. The process of increasing the KB is described in next sub section.

E. Increase the KB

With the association rules approved, the positives ones can be used to update and raise the size of NELL's KB. As not all rules must be generalized, there are two possible ways to increase the KB:

- By populating it with new instances;
- By increasing the relations between domains.

The algorithm populates the knowledge base with new instances by using non generalized association rules [16]. Using the generalized ones, NELL can get relations that were not discovered before by its algorithms [17]. Suppose that our approach extracted the following rule:

AthletePlaysInStadium(X, Y) → *AthletePlaysSport(X,Z).*

So, a new relation is discovered. Originally, the KB was consisted only with those two relations *(AthletePlaysInStadium(X, Y)* and AthletePlaysSport(X,Z)) separately. In this way, we can extend NELL's KB by increasing it with a new relation as is illustrated by Figure 2.

IV. Experiments and Results

The experiments taken in this paper have the following main goals:

Table 3: Example of instances from players, sports and championships

Player	Sport	SportsLeague
Kaka	Soccer	Italian_championship
Pirlo	Soccer	Italian_championship
Messi	Soccer	
Hernanes	Soccer	Italian_championship
Cavani	Soccer	
Gilardino		Italian_championship
Cristiano Ronaldo	Soccer	

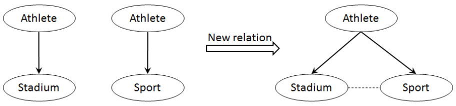

Figure. 2: Example of new relation

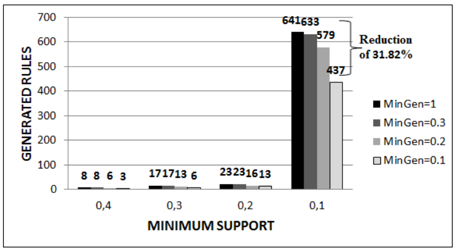

Figure. 4: Extracted rules with generalization process

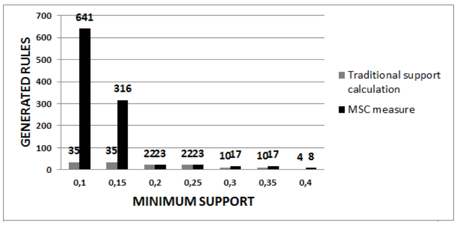

Figure. 3: Extracted rules using MSC measure x traditional approaches

- How association rules may help populating a large KB;
- How the new MSC measure works in a large growing knowledge base consisted of many missing values;
- Verify how generalized association rules can increase a large growing knowledge base like NELL;
- How the techniques developed to deal with obvious rules and itemsets impact and simplify the evaluating process.

The NELL's KB subset used in these experiments has a sports domain, containing eight categories: *athlete, coach, league, stadium, sport, injured, awardTrophyTournament and sportsTeam.*

First experiment has the goal to check rules extracted without MSC measure being applied comparing to our approach. We want to analyze if this new parameter bring more meaningful association rules and how it affects the process in a large growing knowledge base. We did not consider generalization and obvious itemsets and rules treatment in this experiment.

The dataset contains many missing values. As aforementioned, there are two causes for the missing values: instances are not available yet or instances might not exist. Figure 3 shows the amount of rules extracted by each technique.

We performed this test using different minimum support values, varying from 0.04 (four percent) to 0.015 (one point five percent). These values were selected due the KB characteristic and to investigate how it would impact in the number of discovered rules. We also would like to verify whether different association rules would be discovered, and also, to see which method behaves better with different levels of minimum support. Experiments showed that the number of association rules discovered gets higher as the minimum support value decreases. Running the algorithm with different minimum support values revealed that the number of association rules extracted using MSC measure is bigger than compared with traditional method. Such behavior occurred for all minimum support values tested.

Nevertheless, an interesting fact could be observed. Although our MSC measure approach extracted more association rules than traditional ones, it did not induce some rules that were induced by the last one. There is a particular reason for that behavior: as the new support calculation method makes itemsets support higher, the confidence of each rule decreases. Thus, the new MSC measure brings new rules that were not present in the first traditional association rule mining algorithms. This is result of discarding itemsets having missing values in all domains, as explained before.

The second experiment shows the number of association rules (generalized or not) extracted. We use mingen measure, with different settings (0.1, 0.2, 0.3 and 1), in order to analyze the amount of rules extracted and to check how the generalization process can decrease the effort on evaluating rules. Notice that, as we can check in Figure 4, when mingen value is 0.3 the amount of rules extracted in comparison to the ones extracted having mingen value equal to 1 is pretty much close. Minimum support varies from 0.04 to 0.01. Minimum confidence value is fixed on 0.3. The evaluation effort, in this dataset example, has approximately 31.82% of reduction comparing if we not have used minGen. As minGen gets lower, the number of rules extracted decreases due to generalization and redundancy treatment.

Not all discovered rules are generalized and meaningful. We consider a generalized rule as relevant if all items in the extracted rule are generalized and is validated by Conversing Learning. Also, a rule is relevant if all items are specifics instances. In this way, the KB is populated with new instances between the categories. If a rule has both generalized and specifics items, it cannot always be consider relevant. Some examples of these rules are showed in Table 4.

Table 4 shows some extracted and evaluated association rules. The relevant ones (1, 2 and 6) can be used to extend NELL's KB with new relations (rules 1 and 2) or by

Table 4: Example of analyzed rules after evaluating

Number	Association Rule	Relevant Rule
1	*athletePlaysLeague (X,Y)* → *athletePlaysSport(X,Z)*	Yes
2	*athleteWinAwardTrophyTournament(X,Y)* → *athletePlaysSport(X,Z)*	Yes
3	*athleteWinAwardTrophyTournament(X,super_bowl)* → *athletePlaysLeague(X,Z)*	No conclusion
4	*athletePlaysSport(X,Y)* → *athleteWinAwardTrophyTournament(X,Z)*	No
5	*athleteWinAwardTrophyTournament(X,us_open)* → *athletePlaysSport(X,tennis)*	No conclusion
6	*athletePlaysTeam(X, bulls)* → *athletePlaysSport(X, basketball)*	Yes

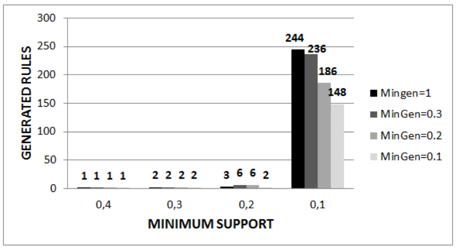

Figure. 5: Extracted rules after applying the techniques

populating it (rule 6). It is noticeable that rule 4 was not consider relevant, once not all players would win a tournament. However, for some association rules (3 and 5), Conversing Learning was not able to give a correct answer. Although rule 3 is a positive one, it brings an association between instances and categories and it is neither possible to update the KB with new instances (since the algorithm does not know the correspondent league) nor with new relations (as the antecedent *super_bowl* is not generalized). Rule 5 describes a homonyms issue since other sports instead of tennis have the same tournament name.

After having properly evaluated the association rules extracted, positive examples are used to update the KB. They are stored and will not be display in future iterations. The same is done to negative examples, as we are not interesting in analyze rules that we already know are irrelevant.

As our approach works on a continuously growing knowledge base, the subset was increased with new facts extracted from web by NELL's algorithm. The subset used got 15% bigger. In order to prove how effective our obvious itemsets and generalized association rules method is, we executed the algorithm again, using same measures values as before.

The number of extracted association rules obtained after rerun the algorithm is showed in Figure 5. The amount of rules to be analyzed reduced considerably, particularly the ones with lower minimum support values. This is the result, mainly, of our methodology to eliminate obvious itemsets and rules and shows the efficacy of our approach.

It is important to relate that new rules, which were not discovered before, were brought by the algorithm as consequence of a large growing knowledge base that, in this experiment, increased approximately in 15%. Some of the new extracted rules are in Table 5.

Making a simple comparison against Figure 4, the number of rules generated reduced a lot. In all cases, the number of rules decreased in more the 50%, getting higher than 90% in some situations, even with new rules been extracted, once the KB was increased.

V. Conclusion

Efforts in large growing knowledge bases have pointing on how to extend them with new instances and relations. Algorithms used to extract new knowledge usually bring too many new facts to be analyzed. In this way, our approach has the following purposes:

- Increase NELL's KB by populating it with instances;
- Increase NELL's KB by adding new relations;
- Decrease the amount of knowledge discovered to simplify the evaluate process.

To accomplish those goals, this article proposed the use of association (generalized association) rules in order to populate NELL's KB with instances or to increase the relations between its categories, and an obvious itemset and generalized association rules treatment. Experiments demonstrated that the former technique can be very useful to populate the KB with instances as well as extend the relations among categories, and also showed that the second one simplifies the effort in analyzing the quantity of rules extracted.

In future works, we aim to explore these issues:

- The use of temporal data in the process of extracting association rules;
- Consider ambiguity and homonyms data in the KB;
- Develop metrics to better evaluate the knowledge obtained;
- Evolve the algorithm to automatically evaluate the rules extracted.

Acknowledgments

I would like to thank Federal Institute of São Paulo (IFSP) and Federal University of São Carlos for financial support and to incentivate research in these organizations.

References

[1] L. A. Galárraga, C. Teflioudi, K. Hose, F. Suchanek. AMIE: Association Rule Mining Under Incomplete Evidence in Ontological Knowledge Bases. In *Proceedings of the 22Nd Int. Conf. on World Wide Web*, pp. 413-422, 2013.

[2] C. Matuszek, J. Cabral, M. Witbrock, J. Deoliveira. An introduction to the syntax and content of Cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, pp. 44-49, 2006.

Table 5: New rules after inserting new facts

Number	Association Rule	Generalized	minGen Values
1	*athletePlaysSportsTeam (X,baltimore_ravens) → athletePlaysLeague(X,nfl)*	No	1.0
2	*athletePlaysSportsTeam (X,la_lakers) → athletePlaysLeague(X,nba)*	No	1.0
3	*athletePlaysSportsTeam(X,knicks) → athletePlaysLeague(X,Z)*	No	0.2
4	*athletePlaysSportsTeam(X,Y) → athletePlaysLeague(X,Z)*	Yes	0.2

[3] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann. DBpedia - A Crystallization Point for the Web of Data, *Journal of Web Semant.*, VII (3), pp. 154-165, 2009.

[4] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, T. M. Mitchell. Toward an architecture for never-ending language learning, *AAAI*, III, pp.3, 2010.

[5] F. M. Suchanek, G. Kasneci, G. Weikum. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pp. 697-706, 2007.

[6] A. Carlson, J. Betteridge, E. R. Hruschka Jr, T. M. Mitchell. Coupling semi-supervised learning of categories and relations, In *Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing*, pp. 1-9, 2009.

[7] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr, T. M. Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web search and data mining*, pp. 101-110, 2010.

[8] A. P. Appel, E. R. Hruschka Jr. In *IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, pp. 917-924, 2011.

[9] T. Mohamed, E. R. Hruschka Jr, T. M. Mitchell. Discovering relations between noun categories. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pp. 1447-1455, 2011.

[10] S. D. S. Pedro, E. R. Hruschka Jr. Conversing learning: Active learning and active social interaction for human supervision in never-ending learning systems. In *Advances in Artificial Intelligence–IBERAMIA*, pp. 231-240, 2012.

[11] J. R. Curran, T. Murphy, B. Scholz. Minimising semantic drift with mutual exclusion bootstrapping. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, 2007.

[12] M. Gardner, P. P. Talukdar, B. Kisiel, T. M. Mitchell. Improving Learning and Inference in a Large Knowledge-Base using Latent Syntactic Cues. In *Conference on Empirical Methods on Natural Language Processing*, pp. 833-838, 2013.

[13] R. Agrawal, T. Imielinski, A. Swami. Mining Association Rules between Sets of Items in Large Databases. In *PROCEEDINGS OF THE 1993 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA*, pp. 207-216, 1993.

[14] R. Srikant, R. Agrawal. Mining Generalized Association Rules. In *Proceedings of the 21th Int. Conf. on Very Large Data Bases*, pp. 407-419, 1995.

[15] R. G. Miani, C. A. Yaguinuma, M. T. P. Santos, M. Biajiz. Narfo algorithm: Mining non-redundant and generalized association rules based on fuzzy ontologies, in *Enterprise Information Systems*, Springer Berlin Heidelberg, pp. 415-426, 2009.

[16] R. G. L. Miani, S. D. S. Pedro, E. R. Hruschla Jr. Association Rules to Help Populating a Never-Ending Growing Knowledge Base. In *Advances in Artificial Intelligence–IBERAMIA 2014*, pp. 169-181, 2014.

[17] R. G. L. Miani, E. R. Hruschla Jr. Analyzing the use of Obvious and Generalized Association Rules in a Large Knowledge Base. In *Proceedings of the 14th International Conference on Hybrid Intelligent Systems*, pp. 1-6, 2014.

[18] R. Navigli, S. P. Ponzetto. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 216-225, 2010.

[19] O. Etzioni, A. Fader, J. Christensen, S. Soderland, Mausam. Open Information Extraction: The Second Generation. In *International Joint Conference on Artificial Intelligence*, pp. 3-10, 2011.

[20] J. Hoffart, F. M. Suchanek, K. Berberich, G. Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia, *Journal of Artificial Intelligence*, 193, pp. 28-61, 2013.

[21] E. Kuzey, G. Weikum. EVIN: Building a Knowledge Base of Events. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pp. 103-106, 2014.

[22] S. Tamang, H. Ji. Relabeling distantly supervised training data for temporal knowledge base population. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pp. 25-30, 2012.

[23] N. Takhirov, F. Duchateau, T. Aalberg. An evidence-based verification approach to extract entities and relations for knowledge base population. In *International Semantic Web Conference*, pp. 575-590, 2012.

[24] W. Shen, J. Wang, P. Luo, M. Wang. Linden: linking named entities with knowledge base via semantic knowledge. In *Proceedings of the 21st international conference on World Wide Web*, pp. 449-458, 2012.

[25] B. Roth, T. Barth, G. Chrupała, M. Gropp, D. Klakow. RelationFactory: A Fast, Modular and Effective System for Knowledge Base Population. In *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 89-92, 2014.

[26] R. G. Miani, C. A. Yaguinuma, Santos, M. T. P, Ferraz, V. R. T. NARFO* Algorithm-Optimizing the Process of Obtaining Non-redundant and Generalized Semantic Association Rules. In *International Conference on Enterprise Information Systems*, pp. 320-325, 2010.

[27] T. Kauppinen, H. Kuittinen, J. Tuominen, K. Seppälä, E. Hyvönen. Extending an ontology by analyzing annotation co-occurrences in a semantic cultural heritage portal. In *Proceedings of the ASWC 2008 Workshop on Collective Intelligence*, pp. 8-11, 2009.

[28] E. Hyvönen, K. Viljanen, J. Tuominen, K. Seppälä. Building a national semantic web ontology and ontology service infrastructure–the FinnONTO approach. Springer Berlin Heidelberg, pp. 95-109, 2008.

[29] N. Nakashole, M. Theobald, G. Weikum. Scalable knowledge harvesting with high precision and high recall. In *Proceedings of the fourth ACM int. conf. on Web search and data mining*, pp. 227-236, 2011.

[30] F. M. Suchanek, M. Sozio, G. Weikum. SOFIE: a self-organizing framework for information extraction. In *SOFIE: a self-organizing framework for information extraction*, 631-640, 2009.

[31] J. R. Nayak, D. J. Cook. Approximate Association Rule Mining. In *FLAIRS Conference*, pp. 259-263, 2001.

[32] A. Rangel, B. Crémilleux. Treatment of missing values for association rules, in *Journal of Research and Development in Knowledge Discovery and Data Mining*, 1394, pp. 258-270, 1998.

[33] M. Hahsler, B. Grün, K. Hornik. A computational environment for mining association rules and frequent item sets. Institut für Statistik und Mathematik, WU Vienna University of Economics and Business, 2005.

[34] T. Calders, B. Goethals, M. Mampaey. Mining itemsets in the presence of missing values. In *Proceedings of the 2007 ACM symposium on Applied computing*, pp. 404-408, 2007.

[35] R. H. Chen, C. M. Fan. Treatment of missing values for association rule-based tool commonality analysis in semiconductor manufacturing. In *IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 886-891, 2012.

[36] T. P. Hong, C. W. Wu. Mining rules from an incomplete dataset with a high missing rate, in *Journal of Expert Systems with Applications*, XXXVIII (4), pp. 3931-3936, 2011.

[37] C. Marinica, F. Guillet. Knowledge-based interactive postmining of association rules using ontologies, in *IEEE Transactions on Knowledge and Data Engineering*, XXII (6), pp. 784-797, 2010.

[38] E. Baralis, L. Cagliero, T. Cerquitelli, P. Garza. Generalized association rule mining with constraints. in *Journal of Information Sciences*, 194, pp. 64-84, 2012.

[39] I. M. A. O. Swesi, A. A. Bakar, A. S. A. Kadir. Mining positive and Negative Association Rules from interesting frequent and infrequent itemsets. In *9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 650-655, 2012.

[40] N. Rai, S. Jain, A. Jain. Mining Interesting Positive and Negative Association Rule Based on Improved Genetic Algorithm(MIPNAR_GA), in *Network and Complex Systems*, III (10), pp. 17-26, 2013.

[41] B. Yang, H. Huang. TOPSIL-Miner: an efficient algorithm for mining top-K significant itemsets over data streams, in *Journal of Knowledge and information systems*, XXIII (2), pp. 225-242, 2010.

[42] G. I. Webb, S. Zhang. Removing trivial associations in association rule discovery. In *Abstracts Published in the Proceedings of the First International NAISO Congress on Autonomous Intelligent Systems (ICAIS 2002)*, 2002.

Author Biographies

Rafael Garcia Leonel Miani Santa Albertina - Brazil, 1984/11/01. PhD Student on Computer Sciences, Federal University of São Carlos - UFSCar, São Carlos, Brazil, 2013 - nowadays. Master of Computer Sciences, Federal University of São Carlos - UFSCar, São Carlos, Brazil, 2009. Bacharel on Computer Sciences, Federal University of São Carlos - UFSCar, São Carlos, Brazil, 2006.

Estevam Rafael Hruschka Junior PhD on Computer Sciences, Federal University of Rio De Janeiro - UFRJ, Rio De Janeiro, Brazil, 2003. Master of Informatic on Brasília University, Brasília, Brazil, 1997. Bacharel on Computer Sciences, Londrina University, Londrina, Brazil, 1994.